

2-23-2018

Authentic research in an introductory geology laboratory and student reflections: Impact on nature of science understanding and science self-efficacy

Elizabeth Moss
Iowa State University

Cinzia Cervato
Iowa State University, cinzia@iastate.edu

Ulrike Genschel
Iowa State University, ulrike@iastate.edu

Lori Ihrig
Follow this and additional works at: https://lib.dr.iastate.edu/ge_at_pubs

 Part of the [Design of Experiments and Sample Surveys Commons](#), [Educational Assessment, Evaluation and Research Commons](#), [Geology Commons](#), [Hydrology Commons](#), and the [Science and Mathematics Education Commons](#)
Iowa State University, egilvie@iastate.edu

The complete bibliographic information for this item can be found at https://lib.dr.iastate.edu/ge_at_pubs/273. For information on how to cite this item, please visit <http://lib.dr.iastate.edu/howtocite.html>.

Authentic research in an introductory geology laboratory and student reflections: Impact on nature of science understanding and science self-efficacy

Abstract

This study examines the effects of an extended authentic research experience on students' understanding of the nature of science (NOS) and self-efficacy toward science. The curriculum of an introductory physical geology lab was transformed to include a six-week, student-driven research project focused on local groundwater and surface water issues. Students' NOS understanding was measured by using a modified version of the Student Understanding of Science and Scientific Inquiry questionnaire ($n = 181$) and their science self-efficacy using a modified vocational self-efficacy survey ($n = 179$). Data were collected on students during four semesters. We found that the combination of having students explicitly reflect on the NOS and working on a research project improves students' overall NOS understanding, and that the increase was higher for female students. We found that for non-STEM students the research project alone had a more positive effect on their NOS understanding. A research project alone did not significantly increase non-STEM students' science self-efficacy, but adding explicit NOS reflections did, and more so for female students. The self-efficacy of STEM students increased more than the one of non-STEM students if they completed both a research project and reflected on NOS concepts. This complex set of results suggests that there are multiple ways to implement an authentic research experience to increase students' NOS understanding, whereas to increase students' self-efficacy, reflections on the NOS are a good strategy, particularly for STEM students. For female students, a promising approach is designing an experiment and reflecting on the NOS.

Keywords

Scientific literacy, nature of science, self-efficacy, gender

Disciplines

Design of Experiments and Sample Surveys | Educational Assessment, Evaluation, and Research | Geology | Hydrology | Science and Mathematics Education

Comments

This is a manuscript of an article published as Moss, Elizabeth, Cinzia Cervato, Ulrike Genschel, Lori Ihrig, and Craig A. Ogilvie. "Authentic research in an introductory geology laboratory and student reflections: Impact on nature of science understanding and science self-efficacy." *Journal of Geoscience Education* 66, no. 2 (2018): 131-146. doi: [10.1080/10899995.2018.1411730](https://doi.org/10.1080/10899995.2018.1411730). Posted with permission.

Authentic research in an introductory geology laboratory and student reflections: Impact on nature of science understanding and science self-efficacy

Running Title: **Authentic research: nature of science understanding and self-efficacy**

Article Type: **Research**

Authors: Elizabeth Moss[1], Cinzia Cervato [2], Ulrike Genschel [3], Lori Ihrig [4], Craig A. Ogilvie [5]

[1] Department of Geological and Atmospheric Sciences, Iowa State University, elizabethlmoss@gmail.com

[2] Department of Geological and Atmospheric Sciences, Iowa State University, cinzia@iastate.edu

[3] Department of Statistics, Iowa State University, ulrike@iastate.edu

[4] Belin-Blank Center for Gifted Education, College of Education, University of Iowa, lori-ihrig@uiowa.edu

[5] Department of Physics and Astronomy, Iowa State University, cogilvie@iastate.edu

Key words: scientific literacy, nature of science, self-efficacy, STEM, gender

ABSTRACT

This study examines the effects of an extended authentic research experience on students' understanding of the nature of science (NOS) and self-efficacy towards science. The curriculum of an introductory physical geology lab was transformed to include a six-week, student-driven research project focused on local groundwater and surface water issues. Students' NOS understanding was measured by using a modified version of the Student Understanding of Science and Scientific Inquiry questionnaire (n=181) and their science self-efficacy using a modified vocational self-efficacy survey (n=179). Data were collected on students during four semesters. We found that the combination of having students explicitly reflect on the NOS and working on a research project improves students' overall NOS understanding, and that the increase was higher for female students. We found that for Non-STEM students the research

project alone had a more positive effect on their NOS understanding. A research project alone did not significantly increase Non-STEM students' science self-efficacy, but adding explicit NOS reflections did, and more so for female students. The self-efficacy of STEM students increased more than the one of Non-STEM students if they completed both a research project and reflected on NOS concepts. This complex set of results suggests that there are multiple ways to implement an authentic research experience to increase students' NOS understanding, while to increase students' self-efficacy, reflections on the NOS are a good strategy, particularly for STEM students. For female students, a promising approach is designing an experiment and reflecting on the NOS.

INTRODUCTION

The President's Council of Advisors on Science and Technology (PCAST) has estimated that to maintain economic competitiveness, one million more undergraduate students must graduate with science, technology, engineering, and math (STEM) degrees than the current graduation rate will produce. Three quarters of this goal could be met by lowering the attrition rate from STEM fields from 60% to 50% (PCAST, 2012). To reduce the current attrition rate we must understand the reasons why these students are leaving STEM fields. Tobias (1990) found that otherwise academically strong undergraduates leave STEM majors because their experience with science courses was only the passive repetition of facts that they did not find intellectually engaging. Seymour and Hewitt (1997) extended these findings by interviewing students who opt out of STEM majors and found that students not only opt out of STEM majors because of a loss of interest in STEM fields, but also due to the belief that a Non-STEM major would offer a better education, as well as poor teaching by STEM faculty. Moreover, female students in particular

leave STEM fields because they do not see the social nature or applicability of scientific careers (Matthews, 1994; Swanbrow, 2005). At the root of many of these issues lies a disconnect between the nature of science (NOS) and how science is taught and represented in the classroom. The NOS describes what science is, how it works, what scientists are like, and what role society plays in influencing science (McComas et al., 1998; Clough, 2007). If students' only experience with science is as a collection of facts to be passively repeated, then they are likely to have a misunderstanding of the role that discovery, invention, imagination, and creativity play in science (Tobias, 1990). Similarly if their experience of science is an isolated endeavor, then students will not learn how and why scientists collaborate. Students' science self-efficacy is an additional key factor in increasing and retaining STEM majors (Lent et al., 1984, 1986; Hackett and Betz, 1989; Pajares and Miller, 1995). Self-efficacy is the belief in one's ability to succeed at a given task (Bandura, 1977). Increasing a student's self-efficacy has the potential to increase their interest in a given career and major (Lent et al., 1994; Luzzo et al., 1999).

Hence, effective instruction in the NOS coupled with authentic experiences within undergraduate science courses has the potential to dispel misconceptions about both science and scientists, increase students' self-efficacy, and thereby increase the retention of STEM majors. As an example, Lopatto (2007) has shown that undergraduate students involved in research experiences are more engaged in science and more likely to pursue a science career.

Postsecondary introductory science courses also offer an opportunity to increase the scientific literacy of the citizenry, because the majority of students in many of these courses are Non-STEM majors. Understanding the NOS underpins what it means to be a scientifically literate

citizen (Shamos, 1995; Rudolph, 2007, National Research Council, 1996; McComas et al., 2000; Holbrook and Rannikmae, 2007; Holbrook and Rannikmae, 2009), i.e. a citizen who can make informed decisions regarding funding for science endeavors, science education, the validity of scientific evidence in the courtroom, and environmental and energy policy decisions. The National Science Board (1996) found that more than 60% of surveyed American adults did not have a basic understanding of how science works. Therefore, correcting misconceptions about the NOS can increase the ability of citizens to make informed decisions and increase scientific literacy. We acknowledge that understanding scientific content is also an important component in scientific literacy (American Association for the Advancement of Science, 1989; National Research Council, 1996; McComas et al., 2000), however changes in students' understanding of course content was not directly assessed in this current study.

In this paper we test four research questions:

1. In an introductory geology laboratory course, is there a different impact on postsecondary students' understanding of the NOS depending on whether students design an authentic research project or conduct the project?
2. In an introductory geology laboratory course, to what extent does having students reflect on the NOS contribute to their understanding of NOS?
3. In an introductory geology laboratory, is there a different impact on postsecondary students' self-efficacy depending on whether students design an authentic research project or conduct the project?

4. Does the increase in NOS understanding and/or science self-efficacy differ for different populations, e.g. STEM vs Non-STEM majors and male vs. female students?

BACKGROUND

Nature of science (NOS)

The NOS describes what science is, how it works, and what scientists are like (McComas et al., 1998; Clough, 2007). Though many different definitions of the “nature of science” exist, there are some agreed-upon statements that describe the NOS; for example, ‘scientific knowledge is tentative’, ‘science requires creativity’, and ‘observations are influenced by prior knowledge and one’s theoretical framework’ (Smith et al., 1997; McComas et al., 1998; Eflin et al., 1999). Lederman (2007) proposed seven key aspects of NOS that expand on these statements: he included that scientific knowledge is empirically based and subjective, socially and culturally embedded, highlighted the distinction between observations and inferences with the latter more impacted by subjective factors, and the function of, and relationships between, theories and laws. Deng et al. (2011) summarized science education literature on students’ view of the NOS. A minority of the studies that they reviewed are based on a unidimensional framework, with students’ view of the NOS placed within a continuum between empiricist (naive) and constructivist (sophisticated) views where scientific knowledge is ‘constructed tentative reality’ (Tsai, 2000). Most of the studies focus on the multidimensional framework of the views of NOS investigating up to 10 key aspects of NOS using both qualitative and quantitative assessment methods (Deng et al., 2011).

As educators, we represent particular aspects of the NOS by the language, activities, and materials employed in the classroom (Robinson, 1969; Carey and Strauss, 1970; Dibbs, 1982; McComas et al., 1998). Through these representations, students develop conceptions of the NOS—many of which are naive and inaccurate (Lederman, 1992; Ryan and Aikenhead, 1992; Clough, 1995a; McComas et al., 1998). Students' misconceptions develop through exposure to misrepresentations or errors present in textbooks, media, scientific papers, and science teaching (Robinson, 1969; Cawthron and Rowell, 1978; Ryan and Aikenhead, 1992; Clough, 1995b; McComas et al., 1998).

Wong and Hodson (2009) pointed out differences between the views held by scientists and those often found in science education textbooks, specifically in textbooks that describe a hierarchical relationship between theories and laws and the view of science as universal, rather than 'tentative'. Differences between the vocabulary used by science instructors and researchers who created the assessment instruments used to evaluate students' view of the NOS can also have a significant impact, and some authors suggest a review of the naive vs. sophisticated construct (e.g., Wong and Hodson, 2009).

In spite of the controversy in science education literature about the meaning of NOS and on how to evaluate students' understanding of it, there is consensus that experiencing the manner in which science is both tentative yet reliable should lead to greater understanding of the nature of science than memorizing the declaration that 'scientific knowledge is tentative' (McComas et al., 2000). Like science content, NOS concepts should be explored, investigated, and wrestled with

to be truly understood, instead of memorized as an additional collection of facts (Clough, 2007, 2011).

Even though science activities based in inquiry often accurately portray the NOS, by themselves they have been shown to be ineffective in changing students' views of the NOS (Lederman, 1992; Abd-El-Khalick and Lederman, 2000; Khishfe and Abd-El-Khalick, 2002). To change students' views of the NOS requires purposefully drawing students' attention to how science and scientists are portrayed in instructional materials and having students reflect on their NOS conceptions as part of the inquiry activities (Abd-El-Khalick et al., 1998; Abd-El-Khalick and Lederman, 2000; Akerson et al., 2000; Khishfe and Abd-El-Khalick, 2002). The explicit NOS reflections that we included in the introductory geology laboratory curricular materials developed for this study are modeled after numerous examples of teaching practices and interventions that have successfully changed students' views on the NOS (Klopfer and Cooley, 1963; Crumb, 1965; Cossman, 1969; Clough 1995a, 1995b).

Self-efficacy

Bandura (1977) developed the concept of self-efficacy, explaining that a person's self-efficacy towards a task is influenced by the following: performance accomplishments, vicarious learning, verbal persuasion, and emotional arousal. The self-efficacy construct of performance accomplishments is of special interest for this study. Performance accomplishments increase self-efficacy when an individual successfully completes a task (Luzzo et al., 1999), and are arguably the most influential factors in changing self-efficacy (Bandura, 1977). Campbell and Hackett (1986) found that college students who successfully completed math problems (a

performance accomplishment) had increased self-efficacy toward math, while students who were unsuccessful at completing the problems had decreased self-efficacy toward math. Similarly Luzzo et al. (1999) found that math self-efficacy and interest in math careers could be increased in students through a performance accomplishment intervention. However, gender plays a difference in the self-efficacy of students enrolled in introductory physics (Nissen and Shamwell, 2016). The extended authentic research project added to the course could serve as a performance accomplishment for students and impact students' science self-efficacy.

Students' attitude toward a discipline can be increased through accomplishments (Freedman, 1997; French and Russell, 2001; Adams et al., 2006; Barbera et al., 2008). Students' attitudes toward science have been shown to serve as a predictor of whether or not a student will continue pursuing more courses in a discipline (Dalgety and Coll, 2006). For example, in introductory chemistry and physics courses, students' attitudes toward science often decreased by the end of the semester for traditional, lecture courses (French and Russell, 2001; Adams et al., 2006; Barbera et al., 2008). In contrast, lab activities and more student-centered teaching strategies have been shown to significantly improve students' attitudes toward science (Freedman, 1997; French and Russell, 2001). Like in this study, the instruments used include questions that address attitudes toward a specific discipline or science in general (Dalgety et al., 2003; Adams et al., 2006).

CONTEXT

Geology 100L is a one-credit introductory lab course associated with the introductory physical geology lecture course offered at a large public U.S. Midwestern research university. Students

enrolled in the lecture are not required to enroll in the lab course; however, the lecture is a prerequisite or co-requisite for the lab. Both the lecture and the lab course can be used as part of the university's general education science requirement. Students in majors that require four or more credits of natural science courses, or are required to take a lab course, enroll in both lecture and lab. Fewer than ten geology majors are enrolled in the lab each semester.

Between three and four sections of the lab course are typically offered during both fall and spring semesters. Each section consists of up to 25 students and meets once a week for two hours. The lab sections are usually taught by graduate student teaching assistants (TAs); however, one TA in the spring 2011 semester was an undergraduate student. Data included in this study were collected during the spring 2011, fall 2011, spring 2012, and fall 2012 semesters. Seven different TAs taught the labs over the four-semester period—three taught only one semester, three taught in two different semesters, and one taught for three of the four semesters.

Only students who completed the pre-test survey, post-test survey, and open-ended questions on the NOS survey or the pre- and post-test survey of the self-efficacy instrument were included in the dataset resulting in 184 students overall (Table 1). Due to three students not completing the NOS survey and five students not completing the self-efficacy survey, we evaluated 181 students' views on the NOS and 179 students' science self-efficacy. Note that when accounting for gender or STEM/Non-STEM major enrollment sample sizes decreased further by at most seven students for which we were missing information on at least one of these variables. Students enrolled in the study represent close to 65% of all the students enrolled in the lab. Approximately 2/3 of the students included in the sample were non-geology and Non-STEM

majors, and 2/3 were freshmen or sophomores. There were slightly more women than men students (54% vs. 46%) in the whole sample, but the two groups were almost equal in all but the spring 2011 semester when women outnumbered men 3:1. We estimate that approximately 10-15% of sample were racially underrepresented students. More men than women were STEM majors (46% vs. 26%). Although we do not have information on the age of the students included in the study, the vast majority of students at this university are traditional and in their late teens to early twenties.

The roles of the authors were: the first author studied the process as part of her MS project, administered the surveys, met weekly with the TA, made regular lab observations, and compiled the results together with the fourth author. The second author was the faculty member overseeing the labs and graduate advisor of the first author. The third author conducted the statistical analyses of the results, and the fifth author is the director of the institutional transformation project. All contributed to the writing of this paper.

EXTENDED AUTHENTIC RESEARCH PROJECT

Embedding research projects within introductory lab courses has the potential to reach a large number of students, and examples are becoming more common in the literature (e.g., National Academies of Sciences, Engineering and Medicine, 2015). There is a broad spectrum of activities that involve students, ranging from ones centered on the process (the scientific process) to the ones focused on the product (e.g., publishable data) (Beckman and Hensel, 2009). On one end of the spectrum, Kortz and van der Hoeven Kraft (2016) described a course-based undergraduate experience where students completed all components of the scientific process:

they designed a research question (“What do college students think about _____?”), gathered data via questionnaires, analyzed their data, drew conclusions from their data, received peer review, and presented their findings via an in-class oral slideshow presentation.

Closer to the product-oriented experience is the Center for Authentic Science Practice in Education’s (CASPiE) approach to course-based authentic research (Weaver et al., 2006), and that was used as a model for the development of the research experience described here; in introductory chemistry labs at Purdue University, students contributed to a real, faculty-led research project, designed their own project or procedure, and do not know the results beforehand. The Freshman Research Initiative (FRI) at the University of Texas, Austin (Cahalan, 2011) has a similar conceptual design.

Drawing from these successful models, the curriculum of Geology 100L was transformed to include an extended authentic research experience through a six-week research project supported by additional inquiry lab activities (Moss and Cervato, 2016).

The six-week research project for Geology 100L focused on groundwater and surface water processes of the local area and was modeled after the work done by Rathburn and Weinberg (2011) at Colorado State University. First, students visited the site to familiarize themselves with the equipment, the wells, and basic groundwater concepts. At this point, students self-assembled in groups of three to four. Each group was tasked with developing two different research questions based on the field site as homework for the week. During the following two lab periods, groups expanded on this initial exploration, and explored their initial research questions

during a ‘mini project’. They did this by developing a conceptual understanding of groundwater flow and hydrology using an ‘ant farm’ model (a cross-section of an aquifer with a series of wells where students can inject dye to simulate point contamination) and a stream table to explore and model the evolution of a stream system (Moss and Cervato, 2016).

To complete the research project, groups of three to four students:

1. Developed research questions and hypotheses about the local water system with topics ranging from the interactions between the surface and groundwater systems to investigating factors that influence water quality;
2. Determined what data to collect to answer their questions (e.g., nitrate concentrations, water levels, water pH);
3. Collected and analyzed data;
4. Summarized their study in a conference-style poster presentation both in class and at an evening poster session, where they interacted with faculty and staff from across campus to discuss their research.

A difference from the CASPiE and FRI models is that our students were not engaged in faculty research projects; instead, their research contributes to a growing database of local water quality data created for this project.

We focused the research project on surface water and groundwater because of the socioscientific issues surrounding these concepts. Knowledge of surface and groundwater has the potential to help students in making effective decisions about how to address water quality issues, such as the environmental impacts of hydraulic fracturing, sources of water, and water shortages. Most

students harbor misconceptions about groundwater that likely exist and persist because of the unseen and abstract nature of groundwater (Dickerson and Dawkins, 2004; Schwartz et al., 2004; Dickerson et al., 2005; Dickerson et al., 2007; Schwartz et al., 2011). Deep understanding of groundwater concepts requires students to use spatial reasoning, which is underdeveloped in some students (e.g., Kali and Orion, 1996; Dickerson et al., 2005). Hence hands-on activities that focus on improving students' spatial reasoning (e.g., puzzles, drawing, mapping) and 3D physical models can help improve students' understanding of groundwater concepts (Baker and Piburn, 1997; Dickerson et al., 2007). Drawing upon this research, before they began working on their research projects, students conducted an investigation using a three-dimensional "ant farm" groundwater model.

The curriculum before and after the transformation is compared in Table 2. Some traditional introductory geology labs were removed to devote time to the extended authentic research project and supporting inquiry labs. The research project was interwoven into the lab curriculum throughout the semester, occupying six of the fifteen weekly lab periods. Moreover, the "traditional" hands-on labs with limited or no student inquiry were modified to increase the role of student inquiry. More detail on the content and level of inquiry in the labs supporting the research project can be found in Moss and Cervato (2016). The phases of the research project are described in the electronic appendix, which includes also the assessment instruments.

REFORMED CURRICULUM IMPLEMENTATION MODELS

The curricular changes were implemented in stages, starting in spring 2011. A benefit of staggering the implementation is that we were able to assess three different implementation

models on students' NOS views in an introductory geology laboratory course: 1) explicit NOS reflection with the design stage of authentic research (DES+REFL), 2) authentic research without explicit NOS reflection (RES), and 3) the combination of authentic research and explicit NOS reflection (RES+REFL). Table 3 summarizes the implementations described below.

Explicit NOS Reflection with the Design Stage of Authentic Research

During the first semester of implementation, spring 2011, the equipment had not yet been purchased and wells were not installed. Teams of students developed research proposals for a project that would study flooding mitigation efforts for a nearby stream. Teams made a poster on their proposal and participated in an evening poster session. During this semester, questions addressing NOS concepts were included on weekly quizzes and students completed a short written reflection about how the experience with the research proposal had influenced their perceptions about the NOS.

Authentic Research Without Explicit NOS Reflection

Following well installation and equipment purchase during the previous summer, the research project was implemented during the fall 2011 and spring 2012 semesters. Since full implementation of the research project increased both TA and student workload, we removed the weekly quizzes and students were not asked to explicitly reflect on NOS ideas in writing. Instead, we used a portion of the weekly TA meeting time to guide TAs in how to lead discussions about the NOS as part of lab activities. However, TAs did not report implementing these NOS discussions with fidelity, and therefore the fall 2011 and spring 2012 semesters included authentic research without NOS reflection.

330

331 **Authentic Research With Explicit NOS Reflection**

332 To address TA implementation issues and provide a more uniform experience for students, we
333 returned to embedding two explicit NOS reflection questions in each week's lab assignment for
334 the fall 2012 semester. These reflection questions were constructed to target the NOS ideas most
335 prominent in each week's lab. For example, after students explored the relationships between
336 plate boundaries, earthquakes, volcanoes, etc. as a part of the *Plate Tectonics* lab (Moss and
337 Cervato, 2016), they were asked: "Think about the overlapping boundaries of two plates that you
338 just worked on (e.g., where the North American and Pacific plate meet). Did each group of
339 students classify these boundaries the same way? Why did groups classify them differently?
340 Why do scientists make different observations and interpretations when looking at the same
341 thing?" Students answered these questions through the online course management system.
342 Students received credit based on the level of detail of their responses and were given feedback
343 by their TAs on any misconceptions that were present in their answers (Kisbaugh et al., 2012).
344 The reflections did not explicitly connect to the students' research projects.

345

346 **ASSESSMENT**

347 **Nature of Science Instrument**

348 To measure students' understanding of the NOS, a modified version of the Student
349 Understanding of Science and Scientific Inquiry (SUSSI) instrument (Liang et al., 2008) was
350 used at the beginning (pre) and end (post) of each semester of implementation (available online
351 as supplement). The SUSSI provides a means to construct more nuanced views of students' NOS
352 understandings than traditional questionnaires as the SUSSI captures both quantitative and

qualitative student responses that can be cross-referenced. This version of the SUSI had 40 questions (32 Likert-scale rating and 8 short answer questions) grouped into eight NOS constructs and provides us with a multi-dimensional view of NOS. The instrument included the five original SUSI constructs from Liang et al. (2008) addressing: (1) observations and inference; (2) tentativeness; (3) social and cultural embeddedness; (4) creativity and imagination; and (5) scientific methods. Three additional constructs were added to the original instrument to address additional key NOS aspects : (1) social interaction among scientists; (2) development/ acceptance of scientific ideas; and (3) discovery and invention in science (Clough et al., 2010; Herman and Clough, 2016). Each construct comprised four statements students responded to with a Likert scale, and one short answer. The SUSI questionnaire was developed for undergraduate students and revised and tested for reliability and validity (Liang et al., 2008). The items on the modified version of the SUSI have been also validated to reliably measure postsecondary students' and secondary science teachers' NOS understandings (Clough et al., 2010; Herman and Clough, 2016). This validation is for a population with low to early engagement in science, i.e. this population interprets the terms within the SUSI instrument as representing their concepts of science. The terms may be understood differently by more experienced STEM students or practising scientists.

In our work, we used Cronbach's alpha to evaluate the internal consistency for this modified version of the SUSI with the population of postsecondary students in this study. Cronbach's alpha values for the instrument as a whole ranged from 0.65 to 0.85 over the four semesters. Posttest alpha values were higher than pretest alpha values for all semesters, with posttest alpha values ranging from 0.70 to 0.84, while pretest alpha values ranged from 0.65 to 0.83. These

values fall within an acceptable range, indicating that the modified version of the SUSSI was consistently reliable throughout the four semesters of study. They are also consistent with values presented in other studies (e.g., Desaulnier Miller et al., 2010).

The language choices used with SUSSI have been discussed by Liang et. al (2008) and by Clough et al. (2010). The word choice that science is simultaneously both reliable and tentative draws on how these terms are used in national and international K-12 science education standards (McComas et al., 2000). An important part of science is the objective role of measurements, tempered by the subjectivity of scientists through decisions on which measurements to make and on the interpretations of these observations. The SUSSI instrument has four questions within this scale; two on subjectivity in selecting measurements (the role of prior knowledge, unbiased scientists), one on observations as facts, and a final question on interpretations of observations. The authors of SUSSI validate that this set of questions can distinguish between students' who hold a view that science is completely objective and those students who understand that subjectivity plays a role. The sub-question (1.c; see Supplementary Materials) on facts contributes less to distinguishing between these two groups (Liang et. al (2008) and Clough et al. (2010)).

The Likert-style responses from the SUSSI were scored on scale of 1 (strongly disagree) to 5 (strongly agree). Some items of the instrument were negatively worded and were reverse-coded prior to the statistical analysis. The overall SUSSI score for each student corresponds to the sum of all scores. The maximum score for the assessment was 160 points (5 points x 32 questions).

The short answer prompt in each category was used to verify the extent to which the quantitative results (Likert-style responses) accurately reflected students' views as expressed in their written responses. Ryan and Aikenhead (1992) found in the development and study of their NOS instrument (the Views on Science-Technology-Society) that when given the option, some (2-9%) students select "I don't understand" or "None of these choices fits my basic viewpoint". With this in mind, the students' qualitative responses were used to establish confidence in the quantitative portion of the assessment. This differs from the approach used by Desaulnier Miller et al. (2010) who scored the open text answers on a scale from 0 (not classifiable) to 3 (informed view) and used them as additional evidence of students' NOS views; we used them to validate the Likert scale responses.

In order to verify agreement between students' qualitative and quantitative responses, three evaluators developed a scoring rubric for the short answer responses by individually evaluating five students' SUSSI assessments from the spring 2011 semester—coding if the Likert scores did, or did not, match student views expressed in the short answer responses. The evaluators discussed any coding disagreements and refined the scoring rubric, evaluating a second set of five student assessments, and established inter-rater reliability at 85%. The remaining 80 student assessments for the spring 2011 semester were divided up, including a five student overlap to verify that the inter-rater reliability levels remained acceptable. In the overlapping portion of the data coded (8 short answers per student for a total of 40 responses for five students), 85% of responses (34/40) were congruently scored by all evaluators. The scoring rubric the evaluators created and used for scoring students qualitative responses is available upon request from the corresponding author.

422

423 For the spring 2011 sample, ninety percent of students' short answer responses (n = 714) were
424 consistent with their Likert responses. This level of consistency between the students' qualitative
425 and quantitative responses builds confidence that students' Likert-style selections accurately
426 reflected the sophistication of their NOS conceptions. Because such a high consistency between
427 short answer and Likert responses was found for spring 2011, the short answer responses for the
428 other semesters were not evaluated.

429

430 **Self-efficacy**

431 To assess students' self-efficacy, the personal efficacy scale of the vocational self-efficacy
432 survey (Riggs et al., 1994) was modified by replacing the words *my job* with the word *science*.
433 For example, an item that originally read "my future in my job is limited because of my lack of
434 skills" was modified to read "my future in science is limited because of my lack of skills."
435 Student interviews were not performed to verify that the students interpreted the word changes as
436 predicted, but Cronbach's alpha for the modified instrument was >0.8 each semester within each
437 pretest and posttest. These results were consistent with other self-efficacy scales (Luzzo et al.,
438 1999), lending confidence that the results from this instrument were reliable.

439

440 The self-efficacy survey was administered at the same time as the SUSSI. Students entered their
441 response for each question in the self-efficacy survey on a scale with 10 representing a favorable
442 response and 1 an unfavorable response. The maximum score for the assessment was 100 points
443 (10 points x 10 questions). The modified version of the self-efficacy instrument used in this
444 study is available online in the Supplemental Material.

445

446 **ASSESSMENT ADMINISTRATION AND ANALYSIS OF RESULTS**

447 The Institutional Review Board reviewed and approved these instruments following federal
448 regulations and the research was determined to be exempt. Each instrument was made available
449 to the students enrolled in Geology 100L through the lab's online course management system
450 during the first two weeks of the semester (referred to as pretest) and again during the last two
451 weeks of the semester (referred to as posttest). Students usually took 20–40 minutes to complete
452 the SUSSI, and 10–15 minutes to complete the self-efficacy survey. Students received course
453 credit for completing each survey: 5 points for each survey, corresponding to approximately 1%
454 of their grade.

455

456 **METHODS**

457 To measure and compare possible treatment effects we utilized the normalized change score c
458 proposed by Marx and Cummings (2007). Unlike merely subtracting pretest scores from posttest
459 scores, normalized change scores capture the change (improvement or retrogression) that
460 occurred over the semester, given the room students had to improve. Thus, subtle changes in
461 pretest and posttest scores may be more clearly seen because small shifts in scores by students
462 with higher pretest scores produce the same relative change as larger improvements in students
463 with lower pretest scores. According to Marx and Cummings (2007) a student's normalized
464 change score is calculated as $c = \text{IF Pre} < \text{Post THEN } (\text{Post} - \text{Pre}) / (\text{Total Score} - \text{Pre}) \text{ ELSE } (\text{Post} - \text{Pre}) / \text{Pre}$. For students with identical pre- and posttest scores $c = 0$, while scores for students were
465 dropped completely from the analysis if these students achieved either 100% or 0%, respectively,
466

of the possible credit on both pre- and post test. Throughout the Results section, all normalized change scores are expressed as a percentage.

We used numerical and graphical summary statistics to describe students' changes in understanding and in self-efficacy. To assess the effect of treatment on a change in students' understanding and on self-efficacy we conducted one- and two-sample t-tests as well as Analysis of Variance (ANOVA). We choose $\alpha=.1$ as the level of significance for all subsequent statistical inference, but encourage the interested reader to consider a statement made by the American Statistical Association (ASA) in 2016 on statistical significance and p-values (Wasserstein and Lazar, 2016). Principle 3 in the statement advises against reducing the findings of a study to the size of a p-value relative to the chosen level of significance. Principle 5 argues that the size of a p-value is not necessarily indicative of effect size but depends on sample size and measurement precision. In light of this statement, we emphasize that given the amount of variability in the data, it is conceivable that a treatment effect might still exist even though the obtained p-value is greater than .1. Because the size of a p-value depends on the sample size, p-values greater but still close to .1 suggest that further studies are required to assess treatment effectiveness as the collected sample might be too small in size to detect a difference.

Assuming that no other internal or extraneous factors had an effect on student learning and self-efficacy, the one-sided one-sample t-test establishes overall effectiveness of treatment, i.e. under the null hypothesis students' understanding does not improve, on average ($\bar{c}=0$) versus the treatment positively affects students' understanding and self-efficacy ($\bar{c}>0$). A one-way ANOVA model was used to globally assess average treatment differences. The ANOVA was then

followed up with pairwise comparisons of treatments using two-sided two-sample t-tests. Lastly, we expanded each one-way ANOVA model to account for students' gender and STEM major choice. Although the normalized change scores are not normally distributed, statistical inference for the mean is still justified due to sufficiently large group sample sizes. A key assumption for mean-based inference is that the sample mean follows at least approximately a Normal distribution, which is guaranteed when the sample comes from a Normal distribution. When sampling from non-Normal distributions as here, the sampling distribution of the sample mean can be sufficiently well approximated for large sample sizes (Mendenhall and Sincich, 2012). A common sample size recommendation is 30, however, satisfactory approximations to the Normal distribution can already be achieved for much smaller sample sizes, about 15, when the underlying distribution is symmetric and without extreme observations (Mendenhall and Sincich, 2012). For the normalized change scores at hand, observed deviations from normality are sufficiently mild that even for smaller sample sizes the resulting sampling distribution of the sample mean approximates the Normal distribution sufficiently well. As a measure of precaution, we also conducted and included the results of permutation tests for the one-way ANOVA analysis and subsequent two-sample treatment comparisons (e.g., Good, 2005). Permutation tests are considered nonparametric tests as they do not require any distributional assumptions to the data. Thus, these procedures provide an alternative to parametric statistical methods (ANOVA or t-tests) when assumptions for these methods are not in place. A drawback of non-parametric methods is that they can be conservative when used under ideal data assumptions. If the Normality assumption were severely violated, then the statistical results based on the parametric methods will typically disagree substantially from the results of the nonparametric methods. Conversely, agreement of the results suggests that the sampling

distribution of the sample mean is sufficiently well approximated and the parametric procedures can be used. Because data were collected at the student level, with students being in the same lab course, an assumption of complete independence among observations may not hold, potentially impacting standard errors in statistical tests. This affects the generalizability of findings to the larger population.

RESULTS

Section Differences

We evaluated results from the SUSSI and self-efficacy instrument results by lab section within each semester to assess the impact of potential lab section effects due to potential extraneous factors and effects due to individual TAs. We found no significant differences.

NOS Understanding

Table 4 provides main summary statistics, for all the students in each treatment, of the normalized change scores shown in percent. We first conducted a one-sided one-sample t-test to establish evidence that, on average, students' understanding increased over the course of the semester regardless of treatment choice (Table 4). We found a significant increase in the average normalized change scores for all three treatments, suggesting an increase in students' NOS understanding, on average. The results for all three treatment groups are given in Table 4. We

also provide 95% confidence intervals for each treatment mean. Because the lower bound of each interval is positive, these results confirm that each of the treatments was productive in increasing students' NOS understanding. Across the three treatments, results from a one-way ANOVA show no statistically significant difference in the means ($F=0.68$, $df= 2, 178$, $p\text{-value}=.5073$, permutation test-based $p\text{-value}=.5023$). This result is supported by the subsequent pairwise treatment comparisons. The combination of an authentic research experience with explicit NOS reflection does not result in a statistically significant different average increase compared to exposing students to only one of the two activities (Table 5). Additionally, no difference is found comparing the authentic research experience only directly with DES+REFL.

Within the SUSSI there are eight separate constructs. In all four of the semesters studied, increases occurred only in students' understanding of the role of imagination and creativity. Three of the four ideas in the imagination and creativity construct saw improvements each semester: "scientists use their imagination and creativity when they analyze and interpret data", "imagination and creativity do not conflict with a need to be unbiased", and "imagination and creativity do not conflict with logical reasoning".

Only one or two questions of other constructs (social and cultural influences, methodology of scientific investigation, and social interaction among scientific researchers) showed a significant change in one or two semesters, insufficient to provide a robust interpretation.

We investigated the effect of all three treatments for different genders by stratifying the data by gender and analyzing the normalized change scores accordingly (Table 6, Figure 1). Based on

the observed means, female students appear to gain more than male students across all three treatments, however, for only one of the differences the gain was statistically significant at $\alpha=.1$. The most significant difference observed was for the treatment that combined explicit weekly NOS reflections with a research project ($t=1.73$, $p\text{-value}=.093$). Note that we defined the difference by subtracting the average gain for male students from the average gain for female students.

To understand the effect of the various implementations on Non-STEM or STEM majors, we disaggregated the normalized change scores and analyzed them by STEM/Non-STEM (Table 7, Figure 2). A significant difference in means between STEM and Non-STEM majors was only observed when students conducted a research project without explicit weekly NOS reflections ($t=-2.29$, $p\text{-value}=.0243$). For the other two treatments, the $p\text{-values}$ are .1356 and .1741, respectively, providing no evidence in support of a true group difference given the data at hand. While tendencies of possible differences can be seen in Figure 2 in general, the amount of variability in the normalized change scores is too large to detect any statistically significant differences.

Self-Efficacy Results

Table 8 contains main summary statistics by treatment of students' normalized change scores of self-efficacy. When students conduct an authentic research project a one-sided one- sample t-test shows that students' self-efficacy does not significantly increase, on average, over the course of the semester (RES: $t=1.04$, $p\text{-value}=.1503$). For the other two treatments, we found a statistically significant increase in students' science self-efficacy when students were engaged in explicit

weekly NOS reflections both with and without the research project (Table 8). The results from the one-way ANOVA analysis support the previous finding ($F=2.38$, $df=2, 176$, $p\text{-value}=.0957$, permutation test-based $p\text{-value}=.1036$). Pairwise comparisons between the treatments (Table 9) suggest that students' self-efficacy might be less affected when students only conduct authentic research projects compared to when they engage in explicit weekly NOS reflections ($t=-2.05$, $p\text{-value}=.0422$).

When we disaggregate the sample by gender (Table 10; Figure 3) our results suggest that explicit weekly NOS reflections positively affect female students' self-efficacy ($t=1.61$, $p\text{-value}=.1142/2$ to adjust for one-sided alternative hypothesis). No statistical difference was found in the means in gains in self-efficacy when the data were disaggregated by STEM vs. Non-STEM (Table 11, Figure 4).

DISCUSSION

Pre- and post-test results of our students' understanding of the NOS show a significant increase in all three treatments. Since NOS was not explicitly included in the lecture, changes in NOS understanding are interpreted as being related to what the students did in the lab. On the other hand, we found no significant difference between treatments where students engaged only in research, worked on NOS reflections and designed a research project, or completed an authentic research with NOS reflections, suggesting that there are multiple ways to implement an authentic research experience to contribute to students' NOS understanding. This has potential ramifications for institutions in their planning and implementation of research-labs, especially

universities or community colleges that may not have access to resources or infrastructure for students to conduct research. However, the limitation of this study is that it has been conducted at one university and the context may matter in the extent of students' gain in their understanding of NOS.

The self-efficacy results suggest an interesting difference in the effect that conducting research has on students: explicit NOS reflections had a significant positive effect on students' science self-efficacy. However, there was no difference between pre- and post-test self-efficacy results when students worked on a research project alone, which was unexpected since one of the driving reasons for engaging students in authentic science research is to make them feel and think more like scientists.

Combining the above results leads to possible recommendations for faculty who are planning to embed research activities into their labs, depending on their goals and target students. If a faculty member wants to increase *both* NOS understanding and self-efficacy of Non-STEM majors, then the faculty member could consider design activities and using these to have students reflect explicitly on the nature of science. This is the treatment that produces gains in efficacy and NOS for Non-STEM majors. This is similar to the findings of Khishfe and Abd-El-Khalick (2002) and Freedman (1997). Alternatively, if target audience is STEM majors, then the faculty member could consider having students conduct research and reflect on their experience. This is the treatment that improves both the NOS understanding and self-efficacy for STEM majors.

Anecdotal student feedback to their TAs is in general positive. Students enjoy spending lab time in the field and learning basic water analysis methods. Most of their interest is in water quality rather than in the physical aspects of groundwater flow.

The overall positive experience of students and TAs with this revised curriculum has made it the standard for this introductory geology lab, and continues to be implemented each semester with minor modifications, including providing the students with a list of possible research questions to choose from and that includes projects that students have worked on in the past.

LIMITATIONS

This study was conducted at a single research-intensive institution and included a limited number of students assessed over the span of four semesters. While all students enrolled in the lab were involved in the research experience, about 65% of them completed all surveys used to assess their NOS understanding and self-efficacy, and might represent a biased subsample of the population. While larger sample sizes would have allowed for a stronger signal in the potential treatment differences, the sample size was sufficiently large to interpret the assessment results for the main groupings (female vs male; STEM vs. Non-STEM); however, it does not allow us to further explore the effect that the interventions had, for example, on female STEM or male STEM students, and limits the extent of the interpretation of our data.

The research design was driven by the learning objectives of the reform rather than an intentional research design plan. Further, the NOS assessment instrument has been validated for students

with little to no science experience and hence there are limitations in the language of the instrument that might have been confusing to STEM majors.

CONCLUSIONS

We have assessed students' Nature of Science (NOS) understanding using a modified version of the Student Understanding of Science and Scientific Inquiry (SUSSI) questionnaire. In this paper we tested four research questions:

1. In an introductory geology laboratory course, is there a different impact on postsecondary students' understanding of the NOS depending on whether students design an authentic research project or conduct the project?
 - a. We found that significant improvements occurred in students' understanding of the NOS in all semesters, but the difference between the three treatments is not significant.
2. In an introductory geology laboratory, To what extent does having students reflect on the NOS contribute to their understanding of NOS?
 - a. Our results suggest that for our participants, a research experience has no significantly different effect than designing research with NOS reflections assigned as homework.
3. In an introductory geology laboratory, is there a different impact on postsecondary students' self-efficacy depending on whether students design an authentic research project or conduct the project?
 - a. Students' self-efficacy toward science showed significant improvements when students explicitly reflected on the NOS, particularly in the semester when

students completed only the design phase of a research project. Increases in students' self-efficacy are encouraging as other studies in the literature report that students' attitudes and self-efficacy toward science decrease as a result of their introductory science courses (French and Russell, 2001; Adams et al., 2006; Barbera et al., 2008). Perhaps students are less likely to feel ineffective with just completing a design where nature or practical constraints do not cause students to doubt their designs, and their experiences are being normalized through their reflections.

- b. If students conduct the authentic research, it seems that students need to complete reflections to achieve gains in self-efficacy through a performance accomplishment. It appears that the research project alone without reflections does not improve students' science self-efficacy.

4. Does the increase in NOS understanding and/or science self-efficacy differ for different populations, e.g. STEM vs Non-STEM majors and male vs. female students?

- a. Female students showed a larger improvement in understanding the nature of science than male students, and more so when they engaged in the research project combined with NOS reflections.
- b. On the other hand the treatment of design + explicit reflection on the NOS had the most impact on female students' self-efficacy.
- c. For Non-STEM majors, conducting a research project without explicit NOS reflection has a larger benefit for NOS understanding than gained by STEM students.

d. On the other hand, STEM students engaged in a research project with NOS reflections showed a larger increase in science self-efficacy. It is possible that working on a research project and being given the opportunity to reflect on NOS concepts resonated more positively with students with a declared interest in STEM.

This complex set of results has a few key messages. Firstly, that there are multiple ways to implement an authentic research experience to increase students' NOS understanding. Second, to increase students' self-efficacy, reflections on the NOS seem a good strategy, particularly for STEM students. Finally, for female students a promising approach is designing an experiment and reflecting on the NOS.

ACKNOWLEDGEMENTS

This project was funded by the Howard Hughes Medical Institute. Funding for the field equipment and well installation was provided by the Howard Hughes Medical Institute, the Iowa Math and Science Education Program, the Department of Geological and Atmospheric Sciences and the College of Liberal Arts and Sciences at Iowa State University. We would also like to thank Bill Simpkins, Chris Rehmann, Kristie Franz, Mark Mathison, and Jake Smokovitz for their help with the installation and implementation of the project. We are also grateful to all the geology TAs who taught the labs and patiently assisted us as we transformed the lab curriculum. Constructive criticism from three anonymous reviewers and two associate editors allowed us to significantly improve the original version of this manuscript.

REFERENCES

- Abd-El-Khalick, F. and Lederman, N.G. 2000. Improving Science Teachers' Conceptions of Nature of Science: A Critical Review of the Literature. *International Journal of Science Education*, 22:665–701.
- Abd-El-Khalick, F., Bell, R.L. and Lederman, N.G. 1998. The Nature of Science and Instructional Practice: Making the Unnatural Natural. *Science Education*, 82:417–436.
- Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., and Wieman, C. E. 2006. New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics-Physics Education Research*, 2:010101-1—010101-14.
- Akerson, V.L., Abd-El-Khalick, F. and Lederman, N.G. 2000. Influence of a Reflective Explicit Activity-Based Approach on Elementary Teachers' Conceptions of Nature of Science. *Journal of Research in Science Teaching*, 37:295–317.
- American Association for the Advancement of Science. 1989. Project 2061: Science for all Americans. Washington D.C: AAAS.
- Baker, D. R. and Piburn, M.D. 1997. Constructing science in middle and secondary school classrooms. Boston: Allyn and Bacon.
- Bandura, A. 1977. *Social Learning Theory*. New Jersey: Prentice Hall.
- Barbera, J., Adams, W. K., Wieman, C. E., and Perkins, K. K. 2008. Modifying and Validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry. *Journal of Chemical Education*, 85:1435-1439.
- Beckman, M., and Hensel, N. 2009. Making explicit the implicit: Defining undergraduate research. *Council on Undergraduate Research Quarterly*, 29:40-44.
- Cahalan, R. 2011. Research Program Boosts Student Retention Dramatically. Available at <http://alcalde.texasexes.org/2011/09/research-program-boosts-student-retention-dramatically/> (accessed 22 January 2012).
- Campbell, N. K., and Hackett, G. 1986. The effects of mathematics task performance on math self-efficacy and task interest. *Journal of Vocational Behavior*, 28:149–162.

- Carey, R.L. and Strauss, N.G. 1970. An analysis of experienced science teachers' understanding of the nature of science. *School Science and Mathematics*, 70:366–376.
- Cawthron, E.R. and Rowell, J.A. 1978. Epistemology of science education. *Studies in Science Education*, 7:279–304.
- Clough, M. P., Herman, B. C. and Smith, J. A. R. 2010. Seamlessly teaching science content and the nature of science: Impact of historical short stories on post-secondary biology students. Paper presented at the meeting of the Association for Science Teacher Education (ASTE) National Conference, Sacramento, CA., January 14-16.
- Clough, M.P. 1995a. Longitudinal understanding of the nature of science as facilitated by an introductory high school biology course. In Finley, F., Allchin, D., Rhees, D., and Fitfield, S., eds., *The Proceedings of the Third International History, Philosophy and Science Teaching Conference*, vol 2. Minneapolis, MN: University of Minnesota, p. 212–221.
- Clough, M.P. 1995b. How to accurately portray the nature of science while teaching science content. In Finley, F., Allchin, D., Rhees, D., and Fitfield, S., eds., *The Proceedings of the Third International History, Philosophy and Science Teaching Conference*, vol 2. Minneapolis, MN: University of Minnesota, p. 202–211.
- Clough, M.P. 2007. Teaching the Nature of Science to Secondary and Post-Secondary Students: Questions rather than tenets. *The Pantaneto Forum*, 25.
- Clough, M.P. 2011. Teaching and assessing the nature of science: How to effectively incorporate the nature of science in your classroom. *The Science Teacher*, 78:56–60.
- Cossman, G.W. 1969. The effects of a course in science and culture for secondary school students. *Journal of Research in Science Teaching*, 6:274–283.
- Crumb, G.H. 1965. Understanding of science in high school physics. *Journal of Research in Science Teaching*, 3:246–250.
- Dalgety, J., and Coll, R. K. 2006. Exploring first-year science students' chemistry self-efficacy. *International Journal of Science and Mathematics Education*, 4:97–116.
- Dalgety, J., Coll, R.K. and Jones, A. 2003. The development of the Chemistry Attitudes and Experiences Questionnaire (CAEQ). *Journal of Research in Science Teaching*, 40:649–668.

- Deng, F., Chen, D., Tsai, C.-C., and Chai, C.S. 2011. Students' views of the nature of science: a critical review of research. *Science Education*, 95:961-999.
- Desaulniers Miller, M., Montplaisir, L., Offerdahl, E., Cheng, F., and Ketterling, G. 2010. Comparison of Views of the Nature of Science between Natural Science and Nonscience Majors. *CBE--Life Sciences Education*, 9:45–54.
- Dibbs, D.R. 1982. An investigation into the nature and consequences of teachers' implicit philosophies of science. (Doctoral Dissertation) Birmingham: The University of Aston. Available at http://eprints.aston.ac.uk/13257/1/Dibbs1982_01016112.pdf (accessed 23 October 2011).
- Dickerson, D. L., Callahan, T., Van Sickle, M., and Hay, G. 2005. Students' conceptions of scale regarding groundwater. *Journal of Geoscience Education*, 53:374–380.
- Dickerson, D.L., and Dawkins, K.R. 2004. Eighth grade students' understandings of groundwater. *Journal of Geosciences Education*, 52:178–181.
- Dickerson, D.L., Penick, J.F., Dawkins, K.R., and Van Meter, M. 2007. Groundwater in science education: *Journal of Science Teacher Education*, 18:45–61.
- Driver, R., Leach, J., Scott, P. 1996. Young people's images of science. Bristol, PA: Open University Press.
- Eflin, J.T., Glennan, S., and Reisch, G. 1999. The nature of science: A perspective from the philosophy of science. *Journal of Research in Science Teaching*, 36:107–117.
- Freedman, M. P. 1997. Relationship among laboratory instruction, attitude toward science, and achievement in science knowledge. *Journal of Research in Science Teaching*, 34:343–357.
- French, D.P., and Russell, C.P. 2001. A statistical examination of student achievement and attitude in a large enrollment inquiry-based, introductory, biology course. Paper presented at the meeting of the National Association for Research in Science Teaching. St. Louis, MO. Available at <http://biol1114.okstate.edu/guest/narst-2001.pdf> (accessed 8 June 2012).
- Good, P. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. 3rd Edition. New York, NY, Springer.
- Hackett, G., and Betz, N. E. 1989. An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education*, 20:261–273.

- Herman, B.C. and Clough, M.P. 2016. Teachers' longitudinal NOS understanding after having completed a science teacher education program. *Int. J. of Sci. and Math. Educ.*, 14:207-227. doi:10.1007/s10763-014-9594-1
- Holbrook J. and Rannikmae, M. 2009. The Meaning of Scientific Literacy. *International Journal of Environmental and Science Education*, 4:275–288.
- Holbrook, J., and Rannikmae, M. 2007. Nature of science education for enhancing scientific literacy. *International Journal of Science Education*, 29:1347–1362.
- Kali, Y. and Orion, N. 1996. Spatial abilities of high-school students in the perception of geologic structures. *Journal of Research in Science Teaching*, 33:369-391.
- Khishfe, R. and Abd-El-Khalick, F. 2002. Influence of Explicit and Reflective Versus Implicit Inquiry-Oriented Instruction on Sixth Graders' Views of Nature of Science. *Journal of Research in Science Teaching*, 39:551–578.
- Kisbaugh, T.L.S., Cessnam S., Horst, S.J., Leaman, L., Flanagan, T., Graber Neufeld, D., and Siderhurst, M. 2012. Measuring beyond content: a rubric bank for assessing skills in authentic research assignments in the sciences. *Chemistry Education Research and Practice*, 13:268–276.
- Klopfer, L.E. and Cooley, W.W. 1963. The history of science cases for high schools in the development of student understanding of science: A report of the HOSC instruction project. *Journal of Research in Science Teaching*, 1:33–47.
- Kortz, K.M. and van der Hoeven Kraft, K.J. 2016 Geoscience Education Research Project: Student Benefits and Effective Design of a Course-Based Undergraduate Research Experience. *Journal of Geoscience Education*, 64:24-36.
- Lederman, N.G. 1992. Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching*, 29:331–359.
- Lederman, N.G. 2007. Nature of science: Past, present, and future. In S.K. Abell & N.G. Lederman (eds.). *Handbook of research on science education* (p. 831-879). Mahwah, NJ: Erlbaum.
- Lent, R. W., Brown, S. D., and Hackett, G. 1994. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, 45:79–122.

- Lent, R. W., Brown, S. D., and Larkin, K. C. 1984. Relation of self-efficacy expectations to academic achievement and persistence. *Journal of Counseling Psychology*, 31:356–362.
- Lent, R. W., Brown, S. D., and Larkin, K. C. 1986. Self-efficacy in the prediction of academic performance and perceived career options. *Journal of Counseling Psychology*, 33:165–169.
- Liang, L., Chen, S., Chen, X., Kaya, O., Adams, A., Macklin, M., and Ebenezer, J. 2008. Student Understanding of Science and Scientific Inquiry (SUSSI): Revision and Further Validation of an Assessment Instrument. Paper presented at the National Association for Research in Science Teaching Conference. San Francisco, CA, p. 1-28.
- Lopatto, D. 2007. Undergraduate Research Experiences Support Science Career Decisions and Active Learning. *CBE Life Sci Educ.*, 6:297-306.
- Luzzo, D.A., Hasper, P., Albert, K.A., Bibby, M.A., and Martinelli, E.A. Jr. 1999. Effects of Self-Efficacy-Enhancing Interventions on the Math/Science Self-Efficacy and Career Interests, Goals, and Actions of Career Undecided College Students. *Journal of Counseling Psychology*, 46:233–243.
- Marx, J.D., and Cummings, K. 2007. Normalized change. *American Journal of Physics*, 75:87.
- Matthews, M.R. 1994. History, philosophy, and science teaching: A useful alliance. New York: Routledge.
- McComas, W.F., Clough, M.P., and Almaroza, H. 1998. The role and character of the nature of science in science education. In McComas, W.F., ed., *The Nature of Science in Science Education*. Netherlands: Kluwer Academic Publishers, p. 3–39.
- McComas, W.F., Clough, M.P., Smith, M.U., Lederman, N.G., and Scott, E. 2000. Nature of Science NSTA Position Statement Adopted by the NSTA Board of Directors.
- Mendenhall, W. and Sincich, T. 2012. *A second course in statistics: Regression analysis* (7th ed.). Boston, Mass.; London: Prentice Hall.
- Moss E. and Cervato C. 2016. Quantifying the level of inquiry in a reformed introductory geology lab course. *Journal of Geoscience Education*, 64:125-137.
- National Academies of Sciences, Engineering, and Medicine 2015. *Integrating Discovery-Based Research into the Undergraduate Curriculum*: Report of a Convocation. Washington, DC:

National Academies Press. <http://www.nap.edu/nap-cgi/skimchap.cgi?recid=21851&chap=31-44&act=nap> Accessed on May 26, 2016

National Research Council 1996. National Science Education Standards. Washington, D.C.: National Academy Press.

National Science Board 1996. Science and engineering indicators: 1996. Washington, DC: United States Government Printing Office.

Nissen, J. and Shemwell, J. 2016. Gender, experience, and self-efficacy in introductory physics. *Physical Review Physics Education Research*, 12:020105-1 - 020105-16.

Pajares, F., and Miller, M. D. 1995. Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology*, 42:190–198.

President’s Council of Advisors on Science and Technology 2012. Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President, p. 1-97.

Rathburn, S.L. and Weinberg, A.E. 2011. Undergraduate Student Satisfaction and Achievement at the GetWET Observatory: A Fluid Learning Experience at Colorado State University. *Journal of Geoscience Education*, 59:47–55.

Riggs, M., Warka, J., Babasa, B., Betancourt, R., and Hooker, S. 1994. Development and validation of self-efficacy and outcome expectancy scales for job-related applications. *Educational and Psychological Measurement*, 54:793–802.

Robinson, J.T. 1969. Philosophy of science: Implications for teacher education. *Journal of Research in Science Teaching*, 6:99–104.

Rudolph, J.L. 2007. An inconvenient truth about science education. *Teachers College Record*. Available at <http://www.tcrecord.org> (accessed 23 October 2011).

Ryan, A.G., and Aikenhead, G.S. 1992. Students’ preconceptions about the epistemology of science. *Science Education*, 76:559–580.

Schwartz, K.L., Thomas-Hilburn, H., and Haverland, A. 2011. Grounding Water: Building Conceptual Understanding through Multimodal Assessment. *Journal of Geoscience Education* 59:139–150.

- Schwartz, R.S., Lederman, N.G., Crawford, B.A. 2004. Developing Views of Nature of Science in an Authentic Context: An Explicit Approach to Bridging the Gap Between Nature of Science and Scientific Inquiry. *Science Education*, 88:610–645.
- Seymour, E. and Hewitt, N. 1997. *Talking about Leaving: Why Undergraduates Leave the Sciences*. Boulder, CO: Westview Press.
- Shamos, M. 1995. *The Myth of Scientific Literacy*. New Brunswick: Rutgers University Press.
- Smith, M.U., Lederman, N.G., Bell, R.L, McComas, W.F. and Clough, M.P. 1997. How great is the disagreement about the nature of science? A response to Alters (1997). *Journal of Research in Science Teaching*, 34:1101–1103.
- Swanbrow, D. 2005. *Why Women Shy Away From Careers in Science and Math*. University of Michigan News Service. Available at http://www.ur.umich.edu/0405/Apr11_05/06.shtml (accessed 26 October 2011).
- Tobias, S. 1990. *They're Not Dumb, They're Different: Stalking the Second Tier*. Tucson, AZ: Research Corporation.
- Tsai, C.-C. 2000. Relationship between student scientific epistemological beliefs and perceptions of constructivist learning environments. *Educational Research*, 42, 193–205.
- Wasserstein, R. and Lazar, N. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108
- Weaver, G., Wink, D., Varma-Nelson, P., Lytle, F., Morris, R., Fornes, W., Russell, C., and Boone, W. 2006. Developing a New Model to Provide First and Second-Year Undergraduates with Chemistry Research Experience: Early Findings of the Center for Authentic Science Practice in Education (CASPiE). *The Chemical Educator*, 11:125–129.
- Wong, S. L., and Hodson, D. 2009. From the horse's mouth: what scientists say about scientific investigation and scientific knowledge. *Sci. Educ.* 93:109–130.

999 **TABLES**

1000

1001 **Table 1: Demographic information of 184 students included in this study.** Of these n = 181 completed the SUSSI assessment and
1002 n = 179 completed the self-efficacy survey. We are missing information on gender for n=7 students; for n=3 of these students
1003 information is also missing on STEM/Non-STEM enrollment.
1004 Percentages of the respective populations are included in parentheses.

| Semester | Women STEM | Women Non- STEM | Women Total | Men STEM | Men Non- STEM | Men Total | Gender/ STEM NA | Gender NA | STEM NA | Total |
|------------------------|-----------------------|--------------------------------|------------------------|---------------------|--------------------------|------------------------|--------------------------------|----------------------|--------------------|--------------|
| Spring 2011 | 4 (11.76%) | 30 (88.23%) | 34 (72.34%) | 9 (69.23%) | 4 (30.76%) | 13 (27.66%) | 0 | 0 | 0 | 47 |
| Fall 2011 | 6 (25%) | 18 (75%) | 24 (51.06%) | 8 (36.36%) | 14 (63.64%) | 22 (46.81%) | 0 | 1 (2.13%) | 0 | 47 |
| Spring 2012 | 10 (47.62%) | 11 (5.38%) | 21 (41.18%) | 9 (32.14%) | 19 (67.86%) | 28 (54.90%) | 2 (3.92%) | 0 | 0 | 51 |
| Fall 2012 | 5 (31.25%) | 11 (68.75%) | 16 (41.03%) | 12 (63.16%) | 7 (36.84%) | 19 (48.72%) | 1 (2.56%) | 3 (7.69%) | 0 | 39 |
| Total | 25 | 70 | 95 | 38 | 44 | 82 | 3 | 4 | 0 | 184 |

1005

1006

1007

1008

Table 2: An example of the weekly lab schedule before and after changes to the curriculum in the fall semester. Classes devoted to the research project are in bold font and content that was replaced is shaded.

| | Curriculum | | Research Project |
|------|--|---|---------------------|
| Week | Original | Reformed | Phase |
| 1 | Introduction to measurements and earth processes | Introduction + NOS tubes activity | Pre-project |
| 2 | Plate Tectonics | Introductory Field Activity | Preparation |
| 3 | Earthquakes | Streams and Groundwater (practice investigation) | |
| 4 | Mineral Identification | Streams and Groundwater | |
| 5 | Mineral Identification | Mineral Identification | Research Hypothesis |
| 6 | The Rock Cycle +Igneous Rocks | Rock Identification | |
| 7 | Sedimentary Rocks | Rock Identification | |
| 8 | Metamorphic Rocks | Rock Cycle | Data Collection |

| | | | |
|----|--------------------------------------|-----------------------------------|--|
| 9 | Geologic Time | Field Day | |
| 10 | Stream Processes | Plate Tectonics | |
| 11 | Groundwater Processes | Pangaea | |
| 12 | Geologic Structures and Maps | Work Day | Peer Review and Assessment |
| 13 | Topographic Maps | Topographic Maps | Poster Preparation and Research Presentation |
| 14 | Thanksgiving Break | Thanksgiving Break | |
| 15 | Glacial Processes and Climate Change | Poster Presentations | |
| 16 | Quiz | Geologic Time + Capstone Activity | Post-project |

1009
1010
1011
1012

1013 **Table 3:** Implementation of curricular changes and research design over the four semesters of
 1014 the study.

| | Spring 2011 | Fall 2011 | Spring 2012 | Fall 2012 |
|------------------|-------------|-----------|-------------|-----------|
| Research Project | | x | x | x |
| NOS Reflections | x | | | x |
| Treatment | DES+REFL | RES | RES | RES+REFL |

1015
 1016
 1017
 1018
 1019
 1020
 1021

1022 **Table 4:** Descriptive statistics for the variable normalized change score (SUSI instrument) by
 1023 treatment group and results for one-sided one sample t-test of overall treatment effectiveness.
 1024 Last column contains 95% confidence interval for population level mean normalized change
 1025 score by treatment.

| Treatment | N | Mean | Std. Dev. | Std. Error | t-value | df | p-value | 95% CI |
|-----------|----|--------|-----------|------------|---------|----|---------|--------------------|
| RES | 96 | 7.91% | 22.76% | 2.32% | 3.41 | 95 | .0005 | (3.30%, 12.53%) |
| DES+REFL | 47 | 7.69% | 16.78% | 2.45% | 3.14 | 46 | .0015 | (2.76%, 12.62%) |
| RES+REFL | 38 | 12.33% | 21.57% | 3.50% | 3.53 | 37 | .0006 | (5.25%, 19.42%) |

1026
 1027
 1028
 1029

1030 **Table 5:** Results of two-sided two sample t-test to assess difference in treatment means (SUSSI
1031 instrument) for the variable normalized change score.
1032

| Treatment | Mean Difference | Std. Error | t-value | df | p-value | Permutation based p-value |
|-----------------------|-----------------|------------|---------|-----|---------|---------------------------|
| RES vs. DES+REFL | 0.22% | 3.74% | 0.06 | 141 | .9529 | .9538 |
| RES vs. RES+REFL | -4.42% | 4.30% | -1.03 | 132 | .3057 | .3178 |
| DES+REFL vs. RES+REFL | -4.64% | 4.16% | -1.12 | 83 | .2677 | .2694 |

1033
1034

Table 6: Results of two-sided two sample t-test to assess gender difference in treatment means (SUSSI instrument) for the variable normalized change score within treatment.

| Treatment | Gender | N | Mean | Mean Difference | Std. Error | t-value | df | p-value |
|-----------|--------|----|--------|-----------------|------------|---------|----|---------|
| RES | F | 44 | 10.65% | 5.19% | 4.79% | 1.09 | 91 | .2807 |
| | M | 49 | 5.45% | | | | | |
| DES+REFL | F | 34 | 8.85% | 4.18% | 5.50% | .76 | 45 | .4515 |
| | M | 13 | 4.67% | | | | | |
| RES+REFL | F | 16 | 20.29% | 12.53% | 7.24% | 1.73 | 33 | .0930 |
| | M | 19 | 7.76% | | | | | |

Table 7: Results of two-sided two sample t-test to assess difference in treatment means based on enrollment into STEM (SUSSI instrument) for the variable normalized change score within treatment.

| Treatment | Major | N | Mean | Mean Difference | Std. Error | t-value | df | p-value |
|-----------|----------|----|--------|-----------------|------------|---------|----|---------|
| RES | STEM | 31 | .481% | -11.29% | 4.93% | -2.29 | 92 | .0243 |
| | Non-STEM | 63 | 11.77% | | | | | |
| DES+REFL | STEM | 13 | 2.28% | -7.49% | 5.42% | -1.38 | 45 | .1741 |
| | Non-STEM | 34 | 9.76% | | | | | |
| RES+REFL | STEM | 17 | 18.17% | 10.55% | 6.91% | 1.53 | 33 | .1356 |
| | Non-STEM | 21 | 7.61% | | | | | |

Table 8: Descriptive statistics for the variable normalized change score (Self-efficacy instrument) by treatment group and results for one-sided one sample t-test of overall treatment effectiveness. Last column contains 95% confidence interval for population level mean normalized change score by treatment.

| Treatment | N | Mean | Std.Dev. | Std. Error | t-value | df | p-value | 95% CI |
|-----------|----|--------|----------|------------|---------|----|---------|-----------------|
| RES | 95 | 2.61% | 24.44% | 2.51% | 1.04 | 94 | .1503 | (-2.37%, 7.59%) |
| DES+REFL | 47 | 11.76% | 26.14% | 3.81% | 3.08 | 46 | .0017 | (4.08%, 19.43%) |
| RES+REFL | 37 | 7.55% | 19.55% | 3.21% | 2.35 | 36 | .0122 | (1.04%,14.07%) |

1055 **Table 9:** Results of two-sided two sample t-test to assess difference in treatment means (Self-
 1056 efficacy instrument) for the variable normalized change score.

| Treatment | Mean Difference | Std. Error | t-value | df | p-value | Permutation based p-value |
|----------------------------------|----------------------------|-------------------|----------------|-----------|----------------|--------------------------------------|
| RES vs. DES+REFL | -9.15% | 4.46% | -2.05 | 140 | .0422 | .0357 |
| RES vs. RES+REFL | 4.94% | 4.49% | 1.10 | 130 | .2735 | .2893 |
| DES+REFL vs. RES+REFL | 4.20% | 5.16% | .82 | 82 | .4174 | .4334 |

1057

1058 **Table 10:** Results of two-sided two sample t-test to assess gender difference in treatment means
 1059 (Self-efficacy instrument) for the variable normalized change score within treatment.
 1060

| Treatment | Gender | N | Mean | Mean Difference | Std. Error | t-value | df | p-value |
|-----------|--------|----|--------|-----------------|------------|---------|----|---------|
| RES | F | 45 | 1.24% | -2.37% | 5.10% | -.46 | 90 | .6440 |
| | M | 47 | 3.6% | | | | | |
| DES+REFL | F | 34 | 15.49% | 13.50% | 8.38% | 1.61 | 45 | .1142 |
| | M | 13 | 1.99% | | | | | |
| RES+REFL | F | 16 | 5.26% | -.82% | 5.96% | -.14 | 31 | .8920 |
| | M | 17 | 6.08% | | | | | |

1061

1062 **Table 11:** Results of two-sided two sample t-test to assess difference in treatment means based
 1063 on enrollment into STEM (Self-efficacy instrument) for the variable normalized change score
 1064 within treatment.

| Treatment | Major | N | Mean | Mean Difference | Std. Error | t-value | df | p-value |
|-----------------|----------|----|--------|-----------------|------------|---------|----|---------|
| RES | STEM | 33 | 6.72% | -6.44% | 5.24% | -1.23 | 91 | .2224 |
| | Non-STEM | 60 | .29% | | | | | |
| DES+REFL | STEM | 13 | 13.43% | 2.32% | 8.61% | .27 | 45 | .7891 |
| | Non-STEM | 34 | 11.12% | | | | | |
| RES+REFL | STEM | 17 | 11.16% | 5.87% | 6.55% | .90 | 33 | .3766 |
| | Non-STEM | 19 | 5.29% | | | | | |

1065

FIGURE CAPTIONS

Figure 1: SUSSI normalized change scores (in %) disaggregated by gender. RES: research project only; DES+REFL: Design + NOS reflections; RES+REFL: Research project with NOS reflections.

Figure 2: SUSSI normalized change scores (in %) disaggregated by STEM versus non-STEM majors. RES: research project only; DES+REFL: Design + NOS reflections; RES+REFL: Research project with NOS reflections.

Figure 3: SE normalized change scores (in %) disaggregated by gender. RES: research project only; DES+REFL: Design + NOS reflections; RES+REFL: Research project with NOS reflections.

Figure 4: SE normalized change scores (in %) disaggregated by STEM versus non-STEM majors. RES: research project only; REFL: DES+REFL: Design + NOS reflections; RES+REFL: Research project with NOS reflections.